University of California, Davis

Cognitive Agents for Knowledge Discovery

Award number FA9550-04-1-0159

Final Report: 5 February 2008

P. I: Prof. V Rao Vemuri, UC Davis, rvemuri@ucdavis.edu

1. Statement of Objectives

Discovery of new knowledge, that is, knowledge that we do not already possess, is the focus of this research. This problem can be formulated as an *inverse problem*, where the new knowledge can be represented by the parameters of a black box model. The solution can then be viewed as the culmination of a sequence of problem solving steps: search, composition, integration and discovery. A well designed cognitive agent capable of learning, adaptation and optimization can accomplish this task.

One can seek to automate the entire knowledge discovery process by developing an integrated approach for the search ontology and domain ontology or one can visualize a semi-automated approach where subject matter experts (SMEs) deliberately participate in selected phases of the knowledge discovery process and interact continually with cognitive agents. Both approaches have advantages, depending on the context. In intelligence gathering and analysis, typically one is interested in casting a wide net, gather as much information as possible form a variety of sensors and then make some sense out of it by building models to interpret the data. In domain specific applications, such as Course of Action (COA) applications in Military Decision Making Process (MDMP), the latter approach – although slower and deliberate - gives the SMEs an opportunity to understand knowledge entry, allow knowledge to be collated from different SMEs, and allow knowledge to be validated and maintained in a simplified and efficient manner. Achieving advances in this area is key to providing the information superiority necessary for future DoD mission success. In either case, the underlying knowledge discovery process is essentially the same.

2. Status of Effort

.

- Demonstrated a proof of the concept by formulating the knowledge discovery problem as a machine learning problem wherein the instance-attribute table is assumed to be incomplete, i.e., contains either missing entries or noisy entries. The missing/noisy entries are replenished by searching the WWW for suitable information.
- Developed search and classification methods via the implementation of an on-line, supervised/unsupervised, document clustering technique for web documents. The Naïve Bayes' model was used for supervised document classification and ideas from immune system models were used in the unsupervised mode for document classification.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
05-02-2008	Final	01APR2004 to 1June2007

4. TITLE AND SUBTITLE

Cognitive Agents for Knowledge Discovery

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-04-1-0159

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Professor Vermuri

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

UC Davis

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of
Scientific Research/NL
875 N Randolph St, Ste 325
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-SR-AR-TR-08-0102

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for Public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Discovery of new knowledge, that is, knowledge that we do not already possess, is the focus of this research. This problem can be formulated as an *inverse problem*, where the new knowledge can be represented by the parameters of a black box model. The solution can then be viewed as the culmination of a sequence of problem solving steps: search, composition, integration and discovery. A well designed cognitive agent capable of learning, adaptation and optimization can accomplish this task.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*

- Demonstrated the use of the "aiNet", a hierarchical clustering method, to address complex tasks of document clustering. Based on the immune network and affinity maturation principles, the aiNet is able to remove data redundancy and exhibit good clustering results. Also, Principle Component Analysis (PCA) was integrated into this method to reduce time complexity. The results are compared with HAC and K-means - two classical clustering methods.

3. Accomplishments/New Findings

Summary Description of the Work Performed

The centerpiece of the framework is a cognitive agent (named *Cogent*). The cogent has a built-in capability, at a minimum, for (a) knowledge acquisition, and (b) learning, that is self-adaptive to specific and possibly novel situations. In order to incorporate the variation and evolution motifs, we propose to rely on the *generate-and-test* paradigm wherein different hypotheses are generated in an evolutionary manner by varying a baseline model and testing for validity.

Bayesian Approach

One of the challenges in using Bayesian nets is the need to postulate cause-effect relationships and establish the strength of these relationships by defining conditional probability tables (CPT's) associated with the nodes of the network. Unlike textbook exercises, in real-life situations these data items are very hard to get. One method of getting this information is mining the WWW. One method of validating this information is to use the mined knowledge in an inference engine and compare the inferences drawn from actual experiences. If the inferences derived from these models do not match observed data or subjective experience, the cause-effect relationships implied by the Bayesian nets need adjustment.

To accommodate the subjective elements of the Bayesian approach, an interactive tool for building and modifying the Bayesian nets has been developed. Operationally, the analyst (or user) specifies the initial configuration of the Bayesian net and chooses the corresponding attributes from the given database. In the initial design, the user was given three *options* to specify the dependences among nodes:

- *Naïve Bayes*: the most popular and simplest network structure, given all the nodes.
- *TAN* (Tree Augmented Naïve Bayes, see Figure 1): A method to retrieve a good Bayesian network from training data by searching the space of possible Bayesian networks [Friedman and Goldszmidt, '96].
- *Self-Defined*: This is an option usually for domain experts. They can specify the dependences among the nodes from their experience.

A fourth option, called the *Evolutionary option*, where alternative hypotheses are evolved using ideas from evolutionary computation is yet to be explored.

After the first two steps of building the qualitative part of the Bayesian network, the background inference engine calculates the prior probabilities and CPTs. Once the network is constructed (or evolved) using it solve inference problems is straightforward.

Illustrative Examples Tested

The procedures described above were tested on several data sets: heart disease data, contact lens data, gene expression data and KDD Cup Intrusion Detection data. Detailed results are summarized in the papers listed. Results from other data sets are available in the cited publications. The results show that the method does work and works well.

Relevance to Air Force Mission

Work reported here has many immediate applications to the mission of the Air Force and other services within DoD. For example, work reported here can be used to

(a) Establish practical approaches to simplify the analysis of ever increasing amounts of security–relevant network information already being collected by numerous DoD devices to yield actionable intelligence and situational awareness.

(b) Define secure, innovative new methods for transferring as much—but no more—of the operational data needed to enable effective cooperation between groups that are trying to accomplish a common mission.

(c) Process non-text data sources, such as semi-structured, unstructured images, and spectra using wavelet-based invariant feature extraction techniques.

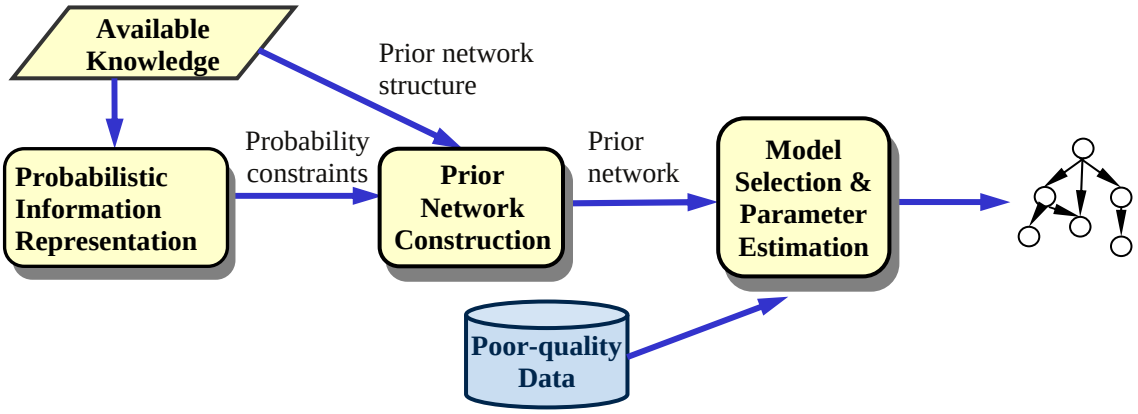

Fig. 1. Framework for Solving the Inverse Problem

5. Personnel Supported

Prof. Rao Vemuri, PI. One month of summer salary
Dr. Na Tang, Doctoral student, Now at Google

6. Publications

1. Vemuri, V. and Tang, N. (2006). “Bayesian Inference by Combining Poor-Quality Data with Knowledge from Subject Matter Experts," Presented at the Software & Systems and Fusion Annual Meeting, 15-19 August 2006, Verona, NY (ppt slides are included in the annual report package)
2. Tang, Na, (2006) “Web Knowledge-based Learning for Ill-posed Inverse Problems,” *Doctoral Dissertation, University of California, Davis, September* (to be submitted).
3. Tang, Na and V Rao Vemuri, (2006) “Bayesian Inference by Combining Poor Quality Data with Knowledge,” *International Journal on Artificial Intelligence Tools*, (submitted).
4. Tang, Na and V. Rao Vemuri, (2005) “An artificial Immune System Approach to Document Clustering,” SAC05, Santa Fe, New Mexico, March 13-15.
5. Tang, Na and V. Rao Vemuri, (2005) “User-Interest-Based Document Filtering via Semi-supervised Clustering,” Lecture Notes in Computer Science, Springer-Verlag, Foundations of Intelligent Systems: *15th International Symposium, ISMIS* Saratoga Springs, NY, USA, May 25-28, 2005. ISBN: 3-540-25878-7
6. Tang, Na and V. Rao Vemuri, (2004) "Web-based Knowledge Acquisition to Impute Missing Values for Classification," *IEEE/WIC/ACM International Joint Conference on Web Intelligence*, Sep. 20-24, Beijing, China.
7. Vemuri, V. and N. Tang, (2004) “Solving Inverse Problems via Machine Learning and Knowledge Discovery," in (Eds. Takumi Ichimura and Katsumi Yoshida.), *Knowledge-Based Intelligent Systems for Healthcare,* CRC Press.
8. Vemuri, V. (2003) “Inverse Problems," in (Eds. G. A. Bekey and B. Y. Kogan.), *Modeling and Simulation: Theory and Practice,* Kluwer Academic Publishers, Boston.

7. Interactions/Transitions

Ideas developed in this research are being used in the ongoing research

- on the design of Next Generation Internet, an NSF project
- on the design of a cyber infrastructure to promote computational thinking in the pursuit of discovery and innovation, an NSF project